УДК 004.91

А.Ю. МИХАЙЛЮК[1], О.В. ПИЛИПЧУК[2], Т.Г. САПСАЙ[2], В.П. ТАРАСЕНКО[2]

[1] ***Київський університет імені Бориса Грінченка, Україна***
[2] ***Національний технічний університет України «Київський політехнічний інститут», Україна***

АВТОМАТИЗОВАНЕ ФОРМУВАННЯ ЛІНГВІСТИЧНОЇ ОНТОЛОГІЇ НА БАЗІ СТРУКТУРОВАНОГО ЕНЦИКЛОПЕДИЧНОГО РЕСУРСУ

Запропонований підхід до вирішення задачі автоматизованого формування лінгвістичної онтологічної бази знань на основі структурованого електронного енциклопедичного ресурсу на прикладі українського сегменту Вікіпедії. Розглядаються основні вимоги щодо формування лінгвістичної онтології та проводиться формалізація процедур конвертації енциклопедичного ресурсу у відповідні елементи онтології із застосуванням теорії графів. Сформована у відповідності до даного підходу онтологія може бути задіяна в задачах підтримки користувача в його інформаційно-пошукової діяльності та інтелектуалізації процедур пошуку, зокрема можливе застосування онтології в процедурах квазісемантичного пошуку.

Ключові слова: *лінгвістична онтологія, структуровані енциклопедичні ресурси, онтологічна база знань, семантичні відношення.*

Вступ

Інтенсивне зростання об'ємів контентного наповнення джерел глобальних та локальних інформаційних баз даних та баз знань актуалізує розробку ефективних інтелектуальних інструментів пошуку інформації, здатних швидко та якісно задовольнити інформаційні потреби користувача. Одним із перспективних напрямків створення таких пошукових інструментів є так званий квазісемантичний пошук [1]. В його основі лежить ідея інтерактивного формування пошукового запиту, як основного джерела інформаційної потреби користувача, за допомогою інтелектуального редактора запиту. Головним засобом реалізації процедур модифікації при формуванні пошукового запиту, згідно концепції квазісемантичного пошуку, є спеціалізований електронний інформаційний ресурс – лінгвістична база знань. Ключовим елементом такого ресурсу є лінгвістична онтологія предметної галузі. Під онтологією розуміють спеціально організовану систему знань, що складається з набору понять і набору тверджень про ці поняття, на основі яких можна будувати внутрішні відношення та зв'язки між поняттями [2].

В контексті квазісемантичного пошуку онтологія на логічному рівні подається семантичною мережею, де спеціальні дескриптори (синсети) утворюють її вузли, пов'язані між собою семантичними відношеннями певних видів. При цьому синсетом називають таку елементарну одиницю онтології, що об'єднує в собі набір синонімів, які описують одне й те саме явище, об'єкт, процес тощо. Таким чином, кожним вузлом подається концепт, що має набір зв'язків-відношень, які з'єднують один вузол-концепт з іншими вузлами-концептами [3]. Зміст концепту відображається його релятивною позицією та унікальним тлумаченням. Кожний концепт має ім'я – символьне позначення, що характеризує значення змісту вузла.

Онтологія, крім того, містить ієрархічну складову, оскільки одним з найбільш важливих видів зв'язків у графі є відношення «елемент класу - клас». Ці зв'язки утворюють онтологічний граф ієрархії концептів природної мови. В корені онтологічної ієрархічної структури знаходиться деякий найбільш абстрактний умовний об'єкт. Від дескрипторної повноти та повноти зв'язків онтології (в тому числі вдалої організації онтологічної ієрархії понять) прямо залежить якість процедур квазісемантичного пошуку. Тому в контексті ефективної реалізації цих процедур надзвичайної актуальності набуває задача створення високоякісної онтологічної бази знань.

На практиці створення лінгвістичної онтології пов'язане з подоланням ряду проблем. По-перше, це складність автоматизації процесу. Кожне поняття в онтології має відповідати певному елементу людського знання. Виділення таких елементів ускладнене нечіткістю змістовних меж об'єктів та процесів реального світу. Труднощі виникають навіть у підготовлених експертів відповідної предметної галузі [4]. Крім того різні експерти цілком правомірно можуть виділяти відмінні набори понять або різну кількість значень для позначення одного і того ж поняття. Очевидно, що автоматизувати процес створення

лінгвістичної онтології фактично дуже складно.

По-друге, крім самих понять, важливим елементом онтологій є семантичні відношення між поняттями в онтології. Інколи зв'язок між поняттями настільки нечіткий або прихований, що однозначно виявити його наявність або кваліфікувати природу (тип відношення) досить важко. Крім того, оцінюючи інтенсивність (вагу) такого зв'язку різні експерти навряд чи будуть одностайні при визначенні його кількісної характеристики.

По-третє, незважаючи на швидкий розвиток глобальної інфраструктури лінгвістичних онтологій (напр., WordNet [5], RusNet [6], EuroWordNet [7] та ін.), для більшості мов світу їх або поки що немає або вони знаходяться в зародковому стані. Така ситуація зокрема характерна і для україномовних онтологій. На перший погляд виходом з даної ситуації міг би бути автоматичний переклад існуючих онтологій іншими мовами. Однак більшість експертів схиляються до думки, що кожна мова відображає свою власну структуру понять та характерні для світосприйняття носіїв даної мови відношення між поняттями. Це, зокрема, засвідчив проект EuroWordNet [7]. Іншими словами, особливості національної культури, побуту, суспільно-політичних відносин тощо можуть суттєво впливати на відображення певних аспектів життєдіяльності в онтології знань і, відповідно, по-різному проектуватись на ту чи іншу мову. Існує й додаткова проблема: автоматизація перекладу надзвичайно ускладнюється тією необхідною умовою, що поняття має перекладатись з урахуванням свого значення, яке залежить від контексту, з яким це поняття взаємодіє. Реалізувати це в автоматичному режимі практично неможливо.

Виходячи із зазначених вище проблем, метою статті є розробка прикладного підходу до автоматизованого формування лінгвістичної онтології, придатної для подальшого використання в засобах інформаційного пошуку, зокрема, в процедурах квазісемантичного пошуку.

1. Постановка задачі автоматизованого формування лінгвістичної онтології

Розглянуті вище проблеми, що виникають при створенні онтологій, а також дослідження існуючих онтологій разом із основними вимогами до цих лінгвістичних ресурсів з боку інформаційно-пошукових систем, дозволяють виділити наступні основні вимоги щодо їх структурно-логічної організації. Вони мають бути максимально враховані при організації процедур формування онтології.

1. Основними інформаційними елементами в онтології мають бути поняття, що характеризуються цілком конкретним значенням і тлумаченням і виражають знання про об'єкт або явище навколишнього світу [8, 9].

2. Кожне поняття повинно інтегрувати в собі всі можливі варіанти лексичного відображення [10]. Іншими словами, необхідно, щоб поняття подавалися синсетами — наборами близьких за змістом синонімів, скорочень тощо.

3. В онтології можуть бути присутні відмінні поняття, що лексично подаються однаково, втім кожне з них має в такому разі характеризуватись різним чітко вказаним тлумаченням [10].

4. Поняття в онтології мають бути пов'язані семантичними зв'язками, що відповідають реальним відношенням між об'єктами, явищами, процесами тощо [8]. Класи відношень, що задіяні в онтології, визначаються вимогами конкретної моделі інформаційно-пошукової системи.

5. Семантичні відношення між поняттями мають фіксуватись або на основі схеми їх зв'язків в рамках даної предметної галузі, яка надається експертами [10, 8], або на основі повного автоматичного семантичного аналізу варіантів вживання понять в предметних текстових корпусах.

6. Необхідно передбачити можливість взаємодії понять онтології з концептами, що виражені іншими мовами [10], для взаємної інтеграції онтологій у багатомовну онтологічну систему і використання в глобалізації процедур пошукових механізмів.

З врахуванням вищенаведених вимог для напівавтоматичного формування лінгвістичної онтології необхідний певний текстовий масив даних. Найчастіше такий масив використовують тематичні колекції публіцистичних видань, підбірки наукових статей або твори художньої літератури. Однак для даного випадку використання таких масивів є проблематичним, оскільки для ідентифікації понять і тим паче зв'язків між ними необхідно проводити складний семантичний аналіз, причому в рамках всього набору текстових документів. Якість такого аналізу на даний час залишається досить низькою. Багатьох подібних труднощів можна уникнути використовуючи як інформаційний масив певні текстові дані Internet-ресурсу енциклопедичного характеру, які уже значною мірою відображають поняття певної предметної галузі та їх взаємовідношення. Очевидним прикладом такої Internet-енциклопедії є Вікіпедія [11]. Зокрема, для створення української онтології можна використати її україномовний сегмент.

Вікіпедія (В) — це відкрита багатомовна енциклопедія знань про оточуючий світ, що колективно створюється користувачами глобальної інформаційної мережі. В її основі лежить принцип колективного наповнення та редагування матеріалів, а також

прийняття чи не прийняття правок на основі загального консенсусу. Важливою особливістю В є безумовно її багатомовність. Таким чином знання відображаються різними мовами світу, в тому числі й українською. Оскільки матеріали В оформлюються у вигляді гіпертексту (тобто в тексті наявні посилання між статтями), можна зробити висновок про те, що В можна вважати прототипом онтології. Таким чином постає задача приведення її до потрібного більш строгого онтологічного формату. Розглянемо детальніше, чому саме такий ресурс може стати основою для формування лінгвістичної онтології і який ефект можна отримати завдяки її використанню в якості інформаційного масиву текстових даних.

1. Кожна стаття В — це фактично опис певного поняття [11, 12], таким чином можна в автоматичному режимі зібрати колекцію понять майбутньої онтології.

2. Кожна (сформована) стаття дає коротке визначення поняття, котре в межах онтології може бути використане як його унікальне тлумачення [13].

3. Однакові за написанням поняття групуються в статті спеціального типу (багатозначне поняття), а кожному з таких понять відповідає окрема сторінка В. Тому, можна вважати, що кожне поняття описується у В з врахуванням його семантики.

4. В описовій частині практично кожної статті В присутні посилання на інші статті [11]. Ці посилання встановлюються вручну при створенні матеріалу авторами статей. Оскільки авторів у такому разі можна вважати експертами в галузі тематики статті, то більшість подібних посилань необхідно трактувати як певний семантичний зв'язок між поняттями, котрі описують ці статті.

5. Кожна стаття і, відповідно, кожне поняття, відносяться до однієї чи більше категорій. Причому кожна категорія сама виступає поняттям і має свою статтю і відповідно свої категорії, підкатегорією котрих вона є. Використовуючи цю властивість, можна побудувати зв'язки типу «гіперонім-гіпонім» між поняттями онтології. Така організація ієрархічних відношень між поняттями буде мати певні особливості порівняно з ієрархією понять в класичних онтологіях, оскільки кожне поняття може бути приписане одразу до декількох категорій [11]. Втім така особливість, на нашу думку, більше відповідає взаємовідношенню понять в реальному світі. Тут лише треба зауважити, що інформаційно-пошукові механізми мають враховувати цю властивість онтології при роботі з нею.

6. Нарешті кожна стаття В містить спеціальні посилання на відповідні статті, подані іншими мовами. При цьому відповідність смислова, тобто статті описують один і той же об'єкт, явище чи процес. Таким чином можна в автоматичному режимі встановлювати іншомовні аналоги понять і зв'язувати онтології сформовані різними мовами, зокрема, наприклад, російською, англійською та іншими. Це, як вже було зазначено, можна використати при розширенні пошукових можливостей інформаційної системи за рахунок пошуку релевантного контенту зарубіжних ресурсів.

Таким чином, наведені властивості статей В та особливості структурної організації самої В, дозволяють запропонувати підхід до формування лінгвістичної онтології на її основі, яка б могла бути використана в роботі інформаційно-пошукових або інформаційно-моніторингових систем.

2. Вікіпедія як прототип лінгвістичної онтології

Для організації автоматизованої обробки ресурсу В розглянемо основні види її статей та структуру останніх. Матеріали всіх статей створюються за допомогою вільного програмного забезпечення MediaWiki [14]. Формат розмітки елементів статті, який пропонується в MediaWiki, набагато простіший і в той же час більш строгий, ніж, наприклад, html-формат. Це дозволяє автоматизувати програмний аналіз структури статті та легко виділяти необхідні її елементи. Для зручності всі матеріали в форматі MediaWiki зібрані в єдиний XML-документ та доступні для завантаження. Окрім безпосередньо вихідного коду статей цей документ містить також різноманітну метаінформацію до кожної з них. Це може бути заголовок, автор, дата тощо. На рис.1 наведено фрагмент XML-документа, що відноситься до однієї із статей, яка присвячена опису поняття «Кортеж». Зокрема для аналізу контента інтерес представляють два поля: поле <title> (заголовок статті) та поле <text> (безпосередньо текст статті).

Усю множину статей В за різними критеріями можна віднести до декількох груп: тлумачні статті, статті, що описують багатозначне поняття, статті категорій, статті, що описують файли, незавершені статті, службові статті. Розглянемо кожну з них.

Тлумачні статті. Основний вид статей, які описують певне поняття, подію або явище. Відповідно до назви є основним елементом інформаційного наповнення енциклопедії та відповідно служать головним джерелом для створення онтології.

Стаття, що описує багатозначне поняття. Спеціальний вид статей, що акумулюють список всіх наявних на даний час у відповідному сегменті В тлумачень конкретного терміну, який заданий в заголовку такої статті. Ці статті містять посилання на відповідну тлумачну статтю по кожному зі значень,

```
<page>
        <title>Кортеж</title>
        <id>19488</id>
        <revision>
                <timestamp>2010-01-21T18:45:56Z</timestamp>
                <contributor><username>Igor Yalovecky</username></contributor>
                <text xml:space="preserve">'''Кортéж''' або "n"-ка – в [[математика|математиці]]
впорядкована та [[скінченна множина|скінченна]] сукупність елементів …
==Дивіться також==
* [[Декартів добуток]]
* [[Формальна мова]]
[[Категорія:Теорія множин]] [[Категорія:Реляційна модель даних]] [[Категорія:Математична нотація]]
[[en:Tuple]][[ru:Кортеж]]
                <text>
        </revision>
</page>
```

Рис. 1. Фрагмент XML-документа Вікіпедії

якщо така є, а також короткий унікальний опис по кожному з семантичних тлумачень терміну. Основний індикатор таких статей – це наявність службової мітки *{{disambig}}* на початку текстової частини. На рис. 2 наведено фрагмент такої статті для терміну "Ядро".

Статті категорій. Статті, що описують поняття-категорію із загальної ієрархії категорій В. В тлумачних статтях одне або декілька таких понять можуть вказуватись як батьківські категорії.

Статті, що описують файли. Крім текстових документів В також містить службові файли, наприклад, зображення. Файлові статті описують специфічні для них дані — посилання, розмір, тип тощо.

Незавершені статті. В тлумачних статтях часто зустрічаються посилання на статті, які ще з тих чи інших причин незавершені. Вони не мають описової частини, але мають заголовок.

Службові статті. Такі статті не несуть безпосереднього інформаційного навантаження і використовуються як додаткові комунікативні та довідкові засоби розробки В. Зокрема це можуть бути шаблони статей, довідки, зауваження, обговорення і т. ін.

В роботі над формуванням онтології на основі статей В ключову роль відіграють тлумачні статті, статті, що описують багатозначне поняття та статті категорій. Причому основу онтології закладають саме група тлумачних статей та група категорій, які можуть бути безпосередньо задіяні у формуванні змістовного наповнення майбутньої онтології. Незважаючи на різницю в назві, статті обох груп дуже близькі за своєю структурою і призначенням, оскільки створюють інформаційний портрет певного поняття. Єдиною суттєвою різницею можна вважати лише те, що статті-категорії описують більш абстрактні або узагальнені поняття і можуть використовуватись для групування підкатегорій або тлумачних статей, тоді як поняття, що описуються в тлумачних статтях, не можуть виступати категоріями для інших об'єктів В. Таким чином, множину статей з цих двох груп можна розглядати як однорідне інформаційне

```
<text xml:space="preserve">{{disambig}}
    *'''Ядро́''' — основна частина, [[Група соціальна|група]] певного [[колектив|колективу]], яка визначає, організує і спрямовує його роботу, діяльність тощо. // [[Основа]] чого-небудь. // у переносному значенні. [[Сутність]], головна [[причина]] чого-небудь.
    *[[Ядро (спорт)]] — спортивне спорядження для легкоатлетичного змагання (штовхання ядра). Переможець визначається за результатом із 6-ти спроб.
    *[[Ядро (атом)]] — центральна частина [[атом|атома]] з розмірами, приблизно у 10 [[мільйон | мільйонів]] разів меншими від розмірів самого атома. Містить [[протон|протони]] і [[нейтрон|нейтрони]].
    *[[Ядро операційної системи]] — фундаментальна частина операційної системи. (Комп'ютерні науки).
    ...
</text>
```

Рис. 2. Фрагмент статті, що описує багатозначне поняття

середовище, що утворює множину інформаційних об'єктів B, а зв'язки між статтями розділити на два класи – звичайні (посилання на інші статті, що присутні безпосередньо в тексті) та категоріальні (посилання, що відображають приналежність статті до тієї чи іншої категорії). Крім того, група статей багатозначних понять виконує допоміжну роль при обробці тлумачних статей по кожному із значень, а незавершені статті можуть бути використанні для ідентифікації зв'язку між іншими статтями, якщо ті будуть містити на неї посилання.

3. Застосування теорії графів для конвертації вікіпедичного ресурсу в онтологічну мережу

Як ресурс документів B, так і онтологія за своєю природою можуть розглядатись у вигляді мережних структур. Тому в рамках підходу до формування онтології для подання цих структур, а також виконання маніпуляцій з її елементами, використаємо апарат теорії графів. При цьому задача формування елементів та зв'язків онтології фактично перетворюється в задачу трансформації (відображення) одного графа в інший за певними алгоритмами. Відповідно до зазначеної теорії, як граф B так і граф онтології будемо описувати двійками $< G, V >$, де G - множина вершин відповідного графа, а V - множина його дуг, що подається підмножиною декартового добутку $G \times G$ [15].

Для початку звернемо увагу на колекцію документів B. Згідно класифікації зв'язків між її статтями загальний граф масиву документів B можна подати як два орієнтовані суграфи (часткові графи). Перший суграф визначає звичайні зв'язки між статтями, а другий - категоріальні. Введемо формальний опис статей B. Позначимо як $D = \{d_1, d_2, ..., d_n\}$ множину документів тлумачних статей (де n - їх кількість), а як $C = \{c_1, c_2, ..., c_m\}$ множину статей-категорій (де m - їх кількість). Загальна множина X інформаційних об'єктів B в такому випадку буде мати вигляд

$$X = \{x_1, x_2, ..., x_{n+m} \mid \forall i \in \overline{1, n+m} : x_i \in D \vee x_i \in C\}.$$

Таким чином, частковий граф документів B, що відображає звичайні посилання між статтями $W^L =< G^X, V^L >$, де G^X - множина вершин, а V^L - множина дуг, можна описати як

$$W^L =< \{x_1, x_2, ..., x_{n+m}\}, \{(x_i, x_j) \mid (x_i \in D \wedge x_j \in D) \vee (x_i \in C \wedge x_j \in D)\} >. \quad (1)$$

Такий запис визначає, що дуги описаного графа можуть сполучати або інформаційні об'єкти, або йти від об'єкта категорії до інформаційного об'єкта. В свою чергу, частковий граф, що відображає категоріальні зв'язки $W^C =< G^X, V^C >$, де V^C - множина дуг, матиме вигляд

$$W^C =< \{x_1, x_2, ..., x_{n+m}\}, \{(x_i, x_j) \mid x_i \in X \wedge x_j \in C\} >. \quad (2)$$

Такий граф міститиме ті дуги, що йдуть від будь-якого об'єкта B, однак лише до об'єктів категорій.

Далі розглянемо лінгвістичну онтологію та її інтерпретацію в рамках теорії графів. Відповідно до [16] формальне подання лінгвістичної онтології можна визначити трійкою $< S, R, \Phi >$, де S - множина синсетів онтології, R - множина семантичних відношень, Φ – множина функцій інтерпретації (в контексті квазісемантичного пошуку – вага семантичних зв'язків між синсетами). Отже множиною вершин графа онтології буде множина синсетів $S = \{s_1, s_2, ..., s_p\}$, де p - кількість синсетів в онтології. Оскільки кожна тлумачна стаття B описує конкретне поняття, то відображення буде прямим: кожній статті відповідатиме синсет онтології. Аналогічно кожна стаття-категорія також прямо відображається у відповідний синсет онтології. Формально це можна записати як

$$\forall d \in D : d \mapsto s, s \in S, \quad \forall c \in C : c \mapsto s, s \in S. \quad (3)$$

Звідси випливає, що $p = n + m$, або кількість синсетів онтології є сумою кількості документів тлумачних статей та кількості статей-категорій. Отже, перше правило відображення графа B у граф онтології буде мати наступний вигляд

$$\forall x_i (x_i \in X, i = \overline{1, n+m}) \; \exists s_i \in S : x_i \mapsto s_i.$$

Тепер розглянемо множину семантичних відношень R онтології. В контексті квазісемантичного пошуку ця множина включатиме в себе два види відношень: відношення гіперонімії r_h та відношення асоціації r_a, тобто $R = \{r_h, r_a\}$. Відповідно до цього граф онтології також можна розділити на два суграфи: частковий граф таксономії понять O^H та частковий граф асоціативних зв'язків онтології O^A. Формально граф $O^H =< G^S, V^H >$, де G^S - множина вершин, а V^H - множина дуг, може бути описаний наступним чином

$$O^H =< \{s_1, s_2, ..., s_p\}, \{(s_i, s_j) \mid s_i \mapsto r_h(w_{ij}) \mapsto s_j\} >, \quad (4)$$

де $r_h(w_{ij})$ - відношення гіперонімії з вагою w_{ij}.

Іншими словами, дуги такого графа будуть об'єднувати ті синсети, між якими є відношення

гіперонімії. Тому граф $O^A =< G^S, V^A >$, де V^A - множина дуг, можна формалізувати як

$$O^A =< \{s_1, s_2, ..., s_p\}, \{(s_i, s_j) | \; s_i \mapsto r_a(w_{ij}) \mapsto s_j\} >, \quad (5)$$

де $r_a(w_{ij})$ - відношення асоціації з вагою w_{ij}.

Тобто дуги графа об'єднують синсети з асоціативним зв'язком. Кожне асоціативне відношення будемо вважати симетричним, тому для кожної пари (s_i, s_j) буде існувати пара (s_j, s_i), причому вага відношення $w_{ij} = w_{ji}$. Таким чином частковий граф асоціативних зв'язків онтології можна вважати неорієнтованим графом.

Відповідно до природи посилань в документах В (звичайні та категоріальні) та природи семантичних відношень онтології (гіперонімії та асоціації) можна поставити відповідність між частковим графом документів Вікіпедії W^C та частковим графом онтології O^H, а також відповідно між підграфами W^L та O^A. Така інтерпретація дозволяє ввести дві операції відображення: відображення звичайних посилань між документами В в асоціативні зв'язки онтології, яке будемо називати конвертацією асоціативних зв'язків, та відображення категоріальних посилань документів В у відповідні зв'язки «гіпонім-гіперонім» онтології, яке будемо називати конвертацією гіперонімічних зв'язків. На рис. 3 та рис.4 зображено графічне подання кожного з наведених видів конвертації для деяких умовних фрагментів відповідних графів.

Введемо поняття ступеня входу по заданій вершині (кількість дуг, що входять у дану вершину із заданої вершини) та ступеня виходу по заданій вершині (кількість дуг, що виходять із даної вершини в задану вершину). Наприклад, на рис. 1 ступінь виходу вершини d_i по вершині d_j дорівнює $\delta^-(d_i, d_j) = 3$, оскільки існує 3 дуги, що йдуть безпосередньо від вершини d_i до вершині d_j. Ступінь входу для вершини d_i по вершині d_j дорівнює $\delta^+(d_i, d_j) = 1$, бо є лише одна дуга, яка йде з вершини d_j у вершину d_i. Очевидно, що вказані характеристики співвідносяться наступним чином: $\delta^+(d_i, d_j) = \delta^-(d_j, d_i)$ та $\delta^-(d_i, d_j) = \delta^+(d_j, d_i)$. Особливість графа В полягає в тому, що зв'язки між документами будуються на основі посилань, які присутні в документах. Таким чином, відповідно до наведених вище визначень, по кожному з документів можна визначити саме ступінь виходу.

Серед особливостей В, що дозволяють використати її для формування онтології, була вказана можливість ідентифікації зв'язків між елементами на основі посилань між статтями. Введені вище характеристики дозволяють сформувати правила відображення зв'язків V^L часткового графа W^L у зв'язки V^A часткового графа онтології O^A

$$\forall (x_i, x_j) \in V^L : \quad (x_i, x_j) \mapsto \{(s_i, s_j), (s_j, s_i)\} \in V^A. \quad (6)$$

При цьому вага відношення залежатиме від кількості прямих та зворотних посилань між двома статтями: чим більше таких посилань наявно в тексті обох статей тим міцнішим має бути зв'язок. Отже,

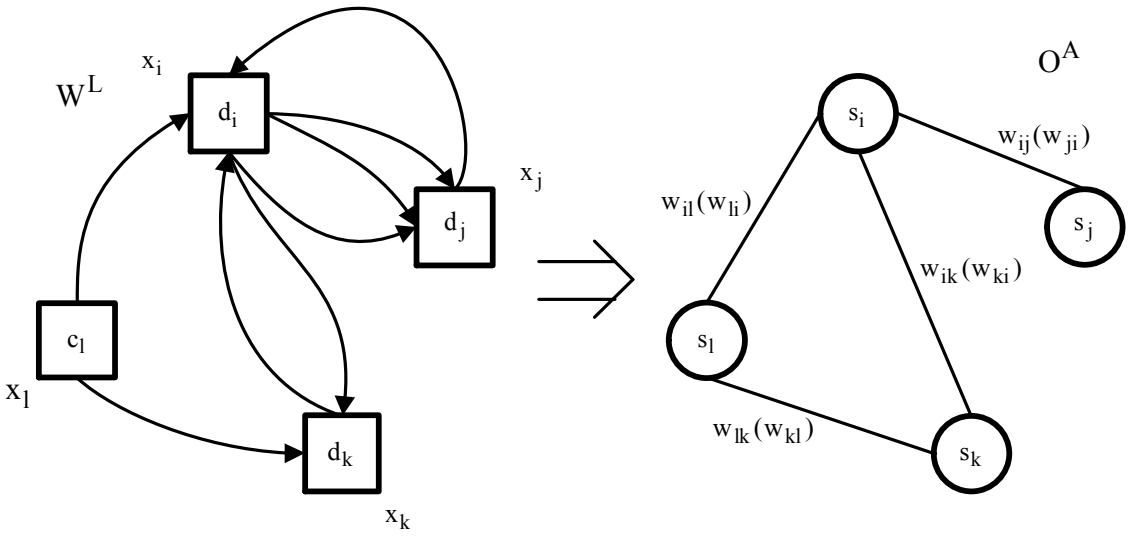

Рис. 3. Конвертація асоціативних зв'язків

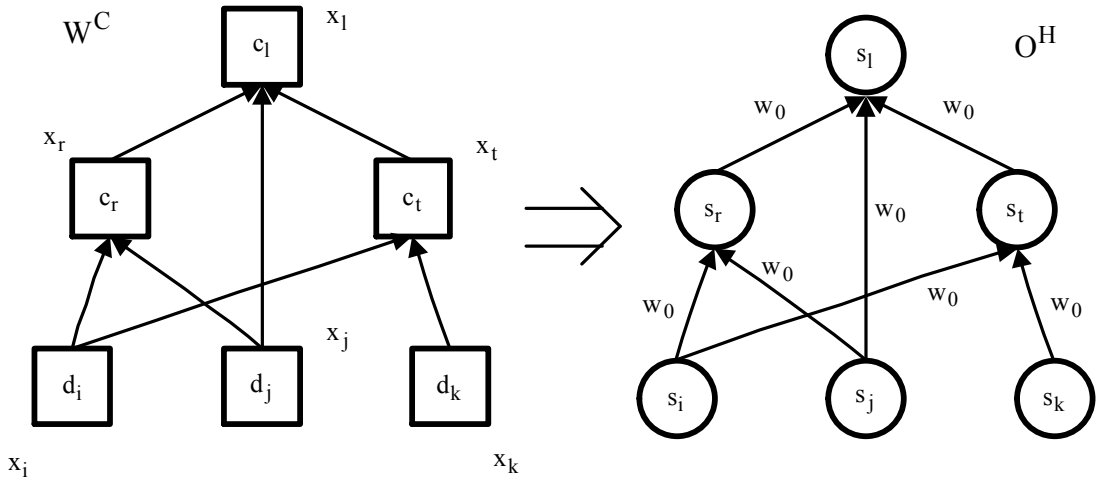

Рис. 4. Конвертація гіперонімічних зв'язків

вага утворених зв'язків (s_i, s_j) та (s_j, s_i) залежатиме від ступенів виходу відповідних вершин початкового графу W^L, а саме: $\delta^-(x_i, x_j)$ та $\delta^-(x_j, x_i)$. Оскільки можливі різні варіанти математичного подання цієї функціональної залежності, розглянемо найпростішу формулу підрахунку ваги відношення синсетів, що забезпечує рівномірність розподілу ваги по онтології:

$$w_{ij} = w_{ji} = w_0 + (\delta^-(x_i, x_j) + \\ +\delta^-(x_j, x_i) - 1) * w_0 * K, \quad (7)$$

де w_0 - деяка одинична вага зв'язку, що відповідає наявності одного посилання між документами B, другий доданок визначає додаткову вагу відношення в залежності від кількості посилань між двома документами, при цьому коефіцієнт K визначає «крутизну» збільшення ваги, підбирається експериментально і лежить в межах $K \in [0,1]$. Завдяки такому визначенню досягається помірна вагова диференціація зв'язків онтології.

У випадку ж відображення зв'язків V^C часткового графу W^C у зв'язки V^H часткового графа онтології O^H правило відображення матиме наступний вигляд

$$\forall (x_i, x_r) \in V^C : (x_i, x_r) \mapsto (s_i, s_r) \in V^H. \quad (8)$$

При цьому, оскільки в частковому графі категорій B можливий лише один зв'язок між двома елементами, то $\delta^-(x_i, x_r) + \delta^-(x_r, x_i) = 1$. Тому вага новоутвореного зв'язку (s_i, s_r) матиме вигляд $w_{ir} = w_0$ і ця вага буде однаковою для всіх дуг часткового графа ієрархії онтології.

Теорема. У результаті конвертації асоціативних та гіперонімічних зв'язків утворюються два часткові графи онтології, множини дуг яких не перетинаються. Або інакше $\forall (s_i, s_j) \in V^A, \neg\exists (s_i, s_j) \in V^H$ та $\forall (s_i, s_j) \in V^H, \neg\exists (s_i, s_j) \in V^A$.

Доведення. Припустимо, що одна і та ж дуга (s_i, s_j) належить одночасно обом частковим графам $(s_i, s_j) \in V^A$ та $(s_i, s_j) \in V^H$. Тоді в першому випадку згідно правила відображення (6) та формального подання (1) існує дуга (x_i, x_j) така, що $x_i \in D \wedge x_j \in D$ або $x_i \in C \wedge x_j \in D$. З цього випливає, що вершина $x_j \in D$. З іншого боку, згідно правила відображення (8) та формального подання (2), існує дуга (x_i, x_j) така, що $x_i \in X \wedge x_j \in C$. З цього випливає, що вершина $x_j \in C$, що суперечить отриманому вище твердженню $x_j \in D$, оскільки множини D та C не перетинаються, тобто $D \cap C = \varnothing$. Таким чином $V^H \cap V^A = \varnothing$, що й треба було довести.

На основі доведеної теореми можна зробити висновок, що частковий граф асоціативних зв'язків та частковий граф ієрархічних зв'язків онтології можна розглядати окремо, а до загального графа онтології входять дуги обох суграфів. А оскільки множини дуг часткових графів онтології не перетинаються, то і вага відношень в об'єднаному графі онтології успадковується від відповідних суграфів без змін.

Висновки

Аналіз проблем, пов'язаних зі створенням онтології (складність автоматизації процесу, нечіткість семантичних відношень, семантична неоднозначність під час перекладу онтології з однієї мови на іншу), показав необхідність пошуку і розробки ефективного підходу до створення таких ресурсів. В контексті сформульованих основних вимог, що мають бути максимально враховані в процесі формування елементів онтологічної бази знань, запропоновано підхід до створення лінгвістичної онтології на основі ресурсу В, а саме: статей В та зв'язків між ними. Інформаційні об'єкти В та гіпертекстові посилання з кожного такого об'єкту на інші можуть розглядатись як своєрідна мережна структура, подібна до мережної структури понять та зв'язків між ними в онтології. Це дало змогу, використовуючи теорію графів, формалізувати основні процедури конвертації ресурсу В у лінгвістичну онтологію. Завдячуючи чіткій структурі матеріалів В, процес формування може бути автоматизований.

Сформована згідно запропонованого підходу онтологія може використовуватись в розробці систем інформаційного пошуку та моніторингу, наприклад, з використанням квазісемантичного підходу до пошуку текстових даних. Іншим напрямком застосування такої онтології може бути дослідження та розробка більш детальних чи спеціалізованих онтологій за участю експертів та інженерів онтологій. Перевага даного підходу полягає у відносній простоті та швидкості процесу автоматичного формування онтологічної бази знань, і, крім того, відкриває потенційно широкі можливості в сфері досліджень багатомовних онтологій.

Таким чином, запропонований у статті підхід до створення лінгвістичної онтології дозволяє у відповідності до наведених процедур конвертації сформувати базову лінгвістичну онтологію, яка включатиме множину понять-синсетів та мережу ієрархічних і асоціативних зв'язків. Зокрема синсети онтології формуються на основі відображення кожної тлумачної статті або статті-категорії у відповідний об'єкт онтології (з необхідними атрибутами, тлумаченням, іншомовними відповідниками тощо), а посилання між статтями В відображаються в мережу семантичних зв'язків між синсетами (з відповідною вагою, що відповідає силі цього зв'язку).

Література

1. Квазісемантичний пошук текстових даних в електронному інформаційному ресурсі [Текст] / А.Ю. Михайлюк, О.В. Пилипчук, М.В Сніжко., В.П. Тарасенко // Радиоэлектроника и інформатика. – Харьков: ХНУРЭ, 2009. – №3. – С. 61-67.

2. Онтологии и тезаурусы [Текст]: Учебное пособие / В.Д. Соловьев, Б.В. Добров, В.В. Иванов, Н.В. Лукашевич. – Казань, Москва, 2006. – 157с.

3. Марченко, О.О. Алгоритми семантичного аналізу природомовних текстів: Автореф. дис. ... канд. фіз.-мат. наук: 05.13.11. – К., 2005. – 17с.

4. Палагин, А.В. Системно-онтологический анализ предметной области [Текст] / А.В. Палагин, Н.Г. Петренко // УСиМ. – 2009. – № 4. – С. 3-14.

5. Fellbaum, C. WordNet: an electronic lexical database [Text] // C. Fellbaum. – MIT Press, 1998. – 423 p.

6. Азарова, И.В. Компьютерный тезаурус русского языка типа WordNet [Текст] / И.В. Азарова, О.А. Митрофанова, А.А. Синопальникова // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции "Диалог–2003". – М.: Издательство РГГУ, 2003. – С. 43-50.

7. Vossen, P. EuroWordNet: a multilingual database with lexical semantic networks [Текст] / P. Vossen. –Dordrecht: Kluwer Academic Publisher, 1998. – 178 p.

8. Абрамов, А.В. Создание лингвистической онтологии образовательной предметной области [Текст] / А.В. Абрамов // Ученые записки. Электронный научный журнал Курского государственного университета. – 2010. – №2. – С. 53-61.

9. Некрашевич, С.П. Построение модели онтологии интеллектуальной системы мониторинга учебного процесса дистанционного образования [Текст] / С.П. Некрашевич // Штучний інтелект. – 2009. – № 2. – С. 124-129.

10. Никоненко, А.А. Обзор баз знаний онтологического типа [Текст] / А.А. Никоненко // Штучний інтелект. – 2009. – № 4. – С. 208-219.

11. Syed, Z.S. Wikipedia as an Ontology for Describing Documents [Text] / Z.S. Syed, T. Finin, A. Joshi / In Proceedings of ICWSM. – AAAI Press, 2008 – P. 136-144.

12. Гладун, А. Семантическая википедия как источник онтологий для интеллектуальных поисковых систем [Текст] / А. Гладун, Ю. Рогушина// Book 2 Advanced Research in Artificial Intelligence. – 2008. – С. 172-178.

13. Kazama, J. Exploiting wikipedia as external knowledge for named entity recognition [Text] / J. Kazama, K. Torisawa // In Proceedings of the 2007 Joint Conference on Computational Natural Language Processing and Computational Natural Language Learning. – 2007. – P. 698-707.

14. Sumida, A. Hacking wikipedia for hyponymy relation acquisition [Text] / A. Sumida, K. Torisawa // In Proceedings of IJCNLP. – 2008. – P. 883-888.

15. Бурков, В.Н. Теория графов в управлении организационными системами [Текст] / В.Н. Бурков, А.Ю. Заложнев, Д.А. Новиков. – М.: Синтег, 2001. – 124 с.

16. Квазисемантический поиск текстовых данных. Способы модификации запроса [Текст] / Д.С. Замятин, А.Ю. Михайлюк, Е.С. Михайлюк, А.В. Петрашенко, А.В. Пилипчук, В.П. Тарасенко // Электронное моделирование. – 2011. – №2. – С. 59-80.

Надійшла до редакції 1.09.2012

Рецензент: д-р техн. наук, професор, професор кафедри автоматизації експериментальных досліджень Є.Т. Володарський, Національний технічний університет України «Київський політехнічний інститут», Київ, Україна.

АВТОМАТИЗИРОВАННОЕ ФОРМИРОВАНИЕ ЛИНГВИСТИЧЕСКОЙ ОНТОЛОГИИ НА ОСНОВЕ СТРУКТУРИРОВАННОГО ЭНЦИКЛОПЕДИЧЕСКОГО РЕСУРСА

А.Ю. Михайлюк, А.В. Пилипчук, Т.Г. Сапсай, В.П. Тарасенко

Предложен подход к решению задачи автоматизированного формирования лингвистической онтологической базы знаний на основе структурированного электронного энциклопедического ресурса на примере украинского сегмента Википедии. Рассматриваются основные требования, касающиеся формирования лингвистической онтологии, и проводится формализация процедур конвертации энциклопедического ресурса в соответствующие элементы онтологии с применением теории графов. Сформированная в соответствии с данным подходом онтология может быть использована в задачах поддержки пользователя в его информационно-поисковой деятельности и интеллектуализации процедур поиска, в частности возможно использование онтологии в процедурах квазисемантического поиска.

Ключевые слова: лингвистическая онтология, структурированные энциклопедические ресурсы, онтологическая база знаний, семантические отношения.

AUTOMATIC CREATION OF LINGUISTIC ONTOLOGY BASED ON A STRUCTURED ENCYCLOPEDIC RESOURCE

A.Y. Mykhailiuk, O.V. Pylypchuk, T.G. Sapsai, V.P. Tarasenko

An approach for solving the task of an automatic creation of a linguistic ontological knowledge base using structured electronic encyclopedic resource (Ukrainian part of Wikipedia as an example) is proposed. The main requirements in creation of linguistic ontology are considered and formalization of converting procedures of encyclopedic resource into appropriate ontology's elements using theory of graphs is carried out. Ontology created in such way can be used in user information retrieval support tasks and at intellectualization of search procedures, in particular it is possible to use created ontology in the procedures of quasi-semantic search.

Key words: linguistic ontology, structured encyclopedic resources, ontological knowledge bases, semantic relations.

Михайлюк Антон Юрійович – канд. техн. наук, старший науковий співробітник, доцент кафедри Інформатики Київського університету імені Бориса Грінченка, Київ, Україна, e-mail: may-62@ukr.net.

Пилипчук Олексій Васильович – асистент кафедри Системного програмування і спеціалізованих комп'ютерних систем факультету Прикладної математики Національного технічного університету України «Київський політехнічний інститут», Київ, Україна, e-mail: ilexcorp@ukr.net.

Сапсай Тетяна Григорівна – канд. техн. наук, доцент кафедри Системного програмування і спеціалізованих комп'ютерних систем факультету Прикладної математики Національного технічного університету України «Київський політехнічний інститут», Київ, Україна, e-mail: stg@scs.ntu-kpi.kiev.ua.

Тарасенко Володимир Петрович – д-р техн. наук, професор, завідувач кафедри Системного програмування і спеціалізованих комп'ютерних систем факультету Прикладної математики Національного технічного університету України «Київський політехнічний інститут», Київ, Україна, e-mail: vtarasen@scs.ntu-kpi.kiev.ua.